A Statistical Rainfall-Runoff Mixture Model with Heavy-Tailed Components

J. Carreau, P. Naveau, Eric Sauquet

Water Resour. Res., 45, W10437, doi:10.1029/2009WR007880.

A Statistical Rainfall-Runoff Mixture Model with Heavy-Tailed Components

J. Carreau[a] P. Naveau[a] E. Sauquet[b]

[a] Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS-UVSQ, Saclay, France

[b] Cemagref Lyon, UR Hydrologie-Hydraulique, Lyon, France

Abstract

We present a conditional density model of river runoff given covariate information which includes precipation at four surrounding stations. The proposed model is non-parametric in the central part of the distribution and relies on Extreme-Value Theory parametric assumptions for the upper tail of the distribution. From the trained conditional density model, we can compute quantiles of various levels. The median can serve to simulate river runoff, quantiles of level 5% and 95% can be used to form a 90% confidence interval, finally, extreme quantiles can estimate the probability of large runoff. The conditional density model is based on a mixture of hybrid Paretos. The hybrid Pareto is built by stitching a truncated Gaussian with a Generalized Pareto distribution. The mixture is made conditional by considering its parameters as functions of covariates. A neural network is used to implement those functions. A penalty term on the tail indexes is added to the conditional log-likelihood to guide the maximum likelihood estimator towards solutions that are preferred. This alleviates the difficulties encounter with the maximum likelihood estimator of the tail index on small training sets. We evaluate the proposed model on rainfall-runoff data from the Orgeval basin in France. The effect of the tail penalty is further illustrated on synthetic data.

1 Introduction

River runoff modelling is relevant for hydroelectricity planning, irrigation and flood prevention. It is a well-known fact among hydrologists that the river runoff is fat-tailed, meaning that sudden large values of runoff can occur which are three or four standard deviations away from the sample mean [BSS+08]. Taking into account those large values is essential since they understandably have a very large impact. Another well-known fact is that precipitation in the hydrographic basin influences the river runoff. However, there are many other mecanisms at work such as underground water tables and soil permeability that are specific to a given hydrographic basin. Most hydrological models try to reproduce the dynamics of the basin by modelling the mecanisms in terms of reservoirs. An alternative approach is to use a stochastic model which provides a full distribution of the river runoff. For example, such a model has been proposed in Lu and Berliner [LB99]. They assume three states or regimes of the runoff process: rising, falling and normal. Transitions probabilities between the states are modelled depending on past runoff values and on rainfall data. Given the current state, the distribution of the river runoff is assumed to follow an autoregressive process which depends on the past runoff values and the observed precipitation. We propose to model the distribution of the runoff at a future time step $t + 1$ given covariate information available at time t with another stochastic model, the conditional mixture of hybrid Paretos presented in [CB08a]. This model bears some similarities to the model of Lu and Berliner [LB99]. In the conditional mixture, we can see the number of components as the number of states, which is determined by model selection instead of being set a priori. The state

selection which is controlled by the mixture weights depends on all the covariates but not on the previous state. The distribution of the river runoff given the current state is given by the corresponding component density, that is a hybrid Pareto density. The parameters of this density are modelled as function of covariates which include past runoff and precipitation. The conditional mixture can adapt to a more general shape of the underlying distribution, including asymmetry and multi-modality. Also, the hybrid Pareto enables the stochastic model to take explicitly extreme values into account. Moreover, a neural network computes, given the covariates, the mixture weights (or state probabilities) and the component density parameters. In contrast to Lu and Berliner [LB99], we don't need to assume a specific form for the relationship between the covariates and the model parameters since such a neural network can in principle approximate any continuous mapping. The model will be further detailed in section 2.

Neural networks have been popular models for a good while in hydrology, see [MD00] for a survey. They were used to predict river runoff but, to our knowledge, not within a conditional mixture framework. Such traditional neural networks are generally not apt at capturing extreme observations. On the other hand, standard models to tackle extremes are drawn from Extreme Value Theory (EVT) [EKM97]. These models consider either maxima over a given period, in which case the generalized extreme-value (GEV) distribution is used, or observations that exceed a selected threshold and a generalized Pareto distribution (GPD) models the distribution of the exceedances. The EVT models thereby mean to estimate the upper tail of the underlying distribution. The choice of the GEV and the GPD is motivated by the fact that these are the limiting distributions of the maxima and

the exceedances respectively under some fairly general conditions. Although extreme runoff behavior is utterly important, hydrologists need to model the whole runoff distribution. One way to extend the GPD model to the whole distribution has been proposed by Frigessi et al. [FHR02]. Their model is a two-component mixture with one light-tailed component and one GPD component. The hybrid Pareto mixture can be seen as a different way to include the GPD into a mixture model. The hybrid is built by stitching together a Gaussian and a GPD while ensuring continuity at the junction point. In the hybrid Pareto mixture, the number of components is chosen according to the data at hand. The central part of the hybrid Pareto mixture consists of a Gaussian mixture which is a flexible non-parametric estimator. The upper tail of the hybrid Pareto mixture is made of a linear combination of GPDs. Through experiments, this approach has shown to perform well on heavy-tailed data [CB08b].

Vrac and Naveau [VN07] have incorporated covariates in the Frigessi mixture [FHR02] in order to predict the distribution of rainfall. The covariates help discriminating between different sorts of rainfall regimes: no rainfall, regular rainfall and extreme rainfall. A particular distribution is used according to which regime prevails. Another way to include covariates into an EVT model has been developed by Chavez-Demoulin and Davison [CDD04]. Covariates are assumed to influence the value taken by the GPD parameters. This relationship is modelled by spline smoothers. In the conditional hybrid Pareto model, the mapping between the hybrid Pareto mixture and the covariates is modelled by a neural network. In this case, the whole conditional distribution is estimated, not just the conditional upper tail, as in the model of Chavez-Demoulin and Davison [CDD04].

The tail index parameter is the most difficult parameter to estimate, whatever model is

used, be it the GPD, the GEV distribution or some other method which one could think of for tail index estimation. This is because the tail index parameter, also termed the shape parameter, gives a sense of the overall shape of the distribution and in particular, of the tail behavior. Typically, few observations will occur in the tail which makes the estimation of the tail index very sensitive. Despite the good asymptotic properties of maximum likelihood estimators (MLEs), they are not very reliable in small samples given their high variance. Estimators of moments show a better behavior in small samples, however they assume that the expectation of the underlying distribution is finite (equivalently, that the tail index is smaller than one). Coles and Dixon [CD99] introduced a penalty term in the MLEs of the GEV parameters. The intuition behind the penalty term is to include a similar range restriction on the tail index estimator as for the moment estimator. Coles and Dixon [CD99] show that the penalized MLE of the tail index performs better in small samples than the classical MLE.

The hybrid Pareto is one such model with a tail index parameter, which is inherited from the GPD. When density estimation is performed with a hybrid Pareto mixture, the tail index of the underlying distribution can be estimated from the tail index of the dominant component in the mixture, that is the component with the largest tail index (and consequently, the heaviest tail). In this case, the MLEs sensitivity in small samples appears in the following way: large tail indexes are assigned to components with negligible mixture weights. To prevent this, we add a penalty term to the log-likelihood based on a prior distribution of the mixture tail indexes. This is similar in spirits to the penalty proposed by Coles and Dixon [CD99]. We devised a prior distribution of the mixture tail indexes based on the following

intuitive idea. We would expect that most components would take care of modelling the central part of the distribution and therefore, have a tail index close to zero. If the tail of the underlying distribution is heavy, we would then expect that some components would have a tail index close to the tail index of the underlying distribution.

We evaluate the conditional hybrid Pareto mixture on rainfall-runoff data from the Orgeval basin in France. The conditional median of the learned conditional hybrid Pareto mixture serves to generate river runoff at a future time step $t+1$. A 90% confidence interval is also computed as the quantiles of level 5% and 95%. This is in contrast with the work of Frigessi et al. [FHR02] and of Vrac and Naveau [VN07] who did not use their model for prediction at a future time step. We also look at the distribution of the conditional tail indexes on the test set; the effect of the tail penalty term in the maximum likelihood estimator can be seen. We gain then more insight into the effect of the new penalty by looking at experiments on synthetic data.

2 Statistical Model of the Rainfall-Runoff Process

We propose to model the rainfall-runoff process with the conditional hybrid Pareto mixture, see [CB08a]. This model combines the flexibility of non-parametric modelling and the extrapolation capability of the GPD methodology. Given a vector of covariates which describe meteorological and hydrological conditions, the conditional distribution of the river runoff is modelled by a mixture of hybrid Paretos whose parameters depend on covariates. Such a mixture is able to adapt to asymmetry, multi-modality and tail heaviness that might be

present in the conditional distribution of the runoff. The neural network which learns the relationship between the covariates and the mixture parameters is able to approximate properly the highly non-linear relationship between rainfall and runoff. The conditional hybrid Pareto mixture provides a conditional density model that has proven to perform well on many kind of data sets (see [CB08a]). The model is explained in details in the following subsections.

2.1 Hybrid Pareto Mixture

Suppose we want to model the distribution of Y, a variable representing the river runoff, with no additional predictive information. We could estimate the distribution of Y with a mixture of Gaussians, which is a popular non-parametric estimator [Bis95]. This type of approach circumvents the need to choose a specific parametric form for the distribution of the runoff and can take into account multi-modality and asymmetry. Mixtures of Gaussians approximate a density by adding up weighted Gaussians or "bumps", see Figure 1. The density estimator is formally given by $\sum_{j=1}^{m} \pi_j \phi_{\mu_j,\sigma_j}(y)$, where the π_j are the mixture weights and $\phi_{\mu_j,\sigma_j}(\cdot)$ is the Gaussian density with parameters μ_j and σ_j. The weights must sum to one, that is $\sum_{j=1}^{m} \pi_j = 1$, to ensure that the estimator is a proper density. A Gaussian mixture approximates the distribution of heavy-tailed data, such as runoff data, by locating one component with a large standard deviation around the largest observations. However, its capacity to extrapolate beyond the sample range might be poor.

The hybrid Pareto distribution was put forward as a way to transfer the extrapolation properties of the GPD [EKM97] to mixture models. The hybrid Pareto distribution is a

smooth extension of the GPD to the whole real axis. This new distribution is built by stitching a GPD tail to a Gaussian, while enforcing continuity of the resulting density and of its derivative. In this work, we focus on runoff data which is heavy-tailed so we let $\xi > 0$ in the GPD density:

$$g_{\xi;\beta}(y-\alpha) = \frac{1}{\beta}(1+\frac{\xi}{\beta}(y-\alpha))^{-1/\xi-1} \quad \xi > 0, \quad y > \alpha.$$

Let α be the junction point and $\phi_{\mu;\sigma}(y) = 1/(\sqrt{2\pi}\sigma)\exp(-(y-\mu)^2/(2\sigma^2))$ be the Gaussian density function with parameters μ and σ. The two constraint equations (equality of the density and of its derivative at α) are solved so that α and β, the GPD scale parameter, become functions of ξ, the GPD tail index and of μ and σ, the Gaussian parameters. Let $\theta = (\xi, \mu, \sigma)$ be the parameter vector of the hybrid Pareto. The hybrid Pareto density is given by:

$$h_\theta(y) = \begin{cases} \frac{1}{\gamma}\phi_{\mu;\sigma}(y) & \text{if } y \leq \alpha, \\ \\ \frac{1}{\gamma}g_{\xi;\beta}(y-\alpha) & \text{if } y > \alpha, \end{cases}$$

where the dependent parameters are $\alpha(\xi,\mu,\sigma) = \mu + \sigma\sqrt{W\left((1+\xi)^2/2\pi\right)}$, $\beta(\xi,\sigma) = (\sigma(1+\xi))/(\sqrt{W\left((1+\xi)^2/2\pi\right)})$ and W is the Lambert W function defined by $w = W(we^w)$ (see [CGH$^+$96]). The re-weighting factor γ ensures that the density integrates to one and is given by:

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + Erf\left(\sqrt{W\left((1+\xi)^2/2\pi\right)/2}\right)\right),$$

where $Erf(\cdot)$ is the error function $Erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt = 2\Phi(z\sqrt{2}) - 1$ and Φ is the standard Gaussian distribution function, (see [PFTV92]). The hybrid Pareto, while inheriting the approximation properties of the GPD, bypasses the need for threshold selection inherent

in the classical GPD methodology [EKM97] since α, the junction point of the Gaussian and the GPD is computed implicitly as a function of the hybrid parameters.

With a hybrid Pareto mixture $\sum_{j=1}^{m} \pi_j h_{\theta_j}(y)$ to model the distribution of the river runoff, we get the best of both worlds: the central part is a mixture of Gaussians which benefits from flexible approximation properties and the upper tail is a linear combination of GPD densities that are capable of extrapolating in areas of unseen data under sound parametric assumptions.

2.2 Conditional Density Model

Our goal is to provide a model of the river runoff at a future time step. We have at our disposal rainfall data in the hydrographic basin of interest which influences river runoff. We therefore look into modelling the distribution of the runoff at time $t + 1$ given covariate information at time t, which includes rainfall observations and past runoff. The hybrid Pareto mixture can be turned into a conditional density model by thinking of the parameters of the mixture as function of covariates [Bis95]. These functions can be implemented in many ways. The simplest model would be a linear model. However, the relationship between rainfall and runoff is highly non-linear. A one-layer feedforward neural network of which the linear model is a special case (no hidden units) is able, if the number of hidden units is well chosen, to approximate any continuous relationship between covariates and mixture parameters. Data-driven selection of the number of hidden units provides a proper level of complexity (or non-linearity). A representation of the conditional mixture model with a neural network is given in Figure 2. The covariates, or inputs, are combined linearly and either fed to the

hidden units or directly connected to the neural network outputs. We took the hyperbolic tangent as the activation function of the hidden layer. The neural network outputs are then transformed into the mixture parameters. Different transformation functions constrain the range of each mixture parameter. The $a_j^{(0)}$ in Figure 2 are dedicated to the mixture weights. The transformation function, the *softmax*, ensures that these weights are positive and sum to one. The $a_j^{(1)}$ and $a_j^{(3)}$ control the tail index and the spread parameter respectively of the j^{th} component. They are guaranteed to be positive by using a *softplus* [DBB$^+$01], a slow-growing version of the exponential. Finally, the $a_j^{(2)}$'s are assigned to the location parameters and need no range constraint.

There are two hyper-parameters to adjust the level of complexity in the conditional hybrid Pareto mixture: the number of hidden units in the neural network and the number of components in the mixture. The former controls the degree of non-linearity of the mapping between the covariates and the mixture parameters and the latter accounts for the complexity of the conditional density (in particular, the multi-modality and asymmetry). Given the approximation capabilities of the neural network and of the mixture model, if the complexity level is well chosen, the conditional mixture should be able to approximate any type of conditional density. The hyper-parameters are chosen so as to maximize the conditional log-likelihood on a validation set, distinct from the training set and thus, should be reasonably close to the ones that give the best generalization performance (the capacity to perform well on unseen data). Because there are many sources of variability (training data, optimization process), the hyper-parameter selection can be variable as well. Overall, the conditional hybrid Pareto mixture gave a better performance than other conditional density estimator

in the presence of heavy-tailed data [CB08a].

2.3 Learning and Regularization

The conditional mixture parameters are the neural network parameters ω. These are learned by minimizing the negative conditional log-likelihood on the training data:

$$\mathcal{L}(\omega) = -\sum_{i=1}^{n} \log(\psi_\omega(y_i|x_i)),$$

where the sum is over the training set $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and $\psi_\omega(y_i|x_i)$ is the hybrid Pareto conditional mixture model evaluated at the data point i.

In [CB08a], the authors have observed empirically that maximum likelihood estimation of the hybrid Pareto mixture, conditional or not, can lead to over-estimation of the tail indexes. This is especially striking for small training sets. The over-estimation of the tail index, even by a small amount, leads to gross over-estimation of the extreme quantiles. In order to guide maximum-likelihood estimation and avoid the over-estimation of the tail indexes, we use a penalty term based on the prior density of Equation (1):

$$f(x; \tau, \eta, \rho) = \tau\eta \exp\{-\eta x\} + (1-\tau)\frac{\exp\{-(x-0.5)^2/(2\rho^2)\}}{\sqrt{2\pi}\rho}. \quad (1)$$

Figure 3 illustrates two typical shapes of the prior density. In the case of runoff data, we can safely assume that the distribution has a tail index around 0.5 ([BSS+08]). This implies that a variant of the full line density in Figure 3 will hold. Most components will be light-tailed, with tail indexes close to zero. These components will take care of modelling the central part of the distribution. Some components will be heavy-tailed, with a tail index value close to the one of the underlying density and these will estimate the upper tail of the

distribution. Hence, the full line density is bimodal, with one mode at zero and the other one, smaller, around 0.5. On the other hand, if the data is light-tailed, then we assume that all the components will have tail indexes close to zero. The prior density in this case would look like the dashed line density in Figure 3.

The two-component mixture of Equation (1) can generate densities such as those illustrated in Figure 3. The exponential component with parameter η controls the density assigned to the small tail indexes and the Gaussian component centered at 0.5 with standard deviation ρ determines how wide the range of the larger tail indexes can be. The mixture weight τ establishes the trade-off between the two components. When τ is equal to zero, we are in the light-tail case.

The conditional mixture parameters ω are now learned by minimizing a new cost function, the negative conditional log-likelihood minus the penalty term:

$$\mathcal{L}(\omega) = -\sum_{i=1}^{n} \log(\psi_\omega(y_i|x_i)) - \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \log f(\xi_{i,j}; \tau, \eta, \rho)$$

where the first sum is over the training set $\mathcal{D}_n$, the second sum in the penalty term is over the number of components m, $\psi_\omega(y_i|x_i)$ is the hybrid Pareto conditional mixture model evaluated at point i and $f(\xi_{i,j}; \tau, \eta, \rho)$ is the prior density evaluated at the tail index of the j^{th} component of the conditional mixture at point i. The penalty term introduces four other hyper-parameters: λ which controls the weight of the penalty with respect to the conditional log-likelihood and τ, η and ρ from the prior density (see Equation (1)). A restricted set of values for the prior density parameters was selected so as to ensure that the prior density follows our prior information about the shape of the distributions of the tail indexes. The

model is trained for several combinations of hyper-parameters (which include the number of hidden units and the number of components of the conditional hybrid Pareto mixture and the hyper-parameters attached to the penalty term). The set of hyper-parameters which gives the smallest cost in terms of negative conditional log-likelihood on data unseen during training (the validation set) is selected.

3 Experiments

We evaluate the conditional hybrid Pareto mixture on the rainfall-runoff data from the Orgeval basin in France. Synthetic data experiments help to gain more insight into the role of the new penalty term in the cost function. Since the generative model is known, the predicted tail indexes can be compared with the tail indexes of the generative model. We also compare the conditional quantiles of the generative versus learned model.

3.1 Orgeval Basin Data

The Orgeval Basin is located in France, East of Paris. There is no snow accumulation in the area that could affect the river runoff. Therefore, we focus on rainfall as a predictor of the river runoff. In order to capture the mecanisms of the basin, moving averages and moving standard deviations of various window lengths of the river runoff are included in the covariates. The river runoff Q_t from the Avenelles sub-basin and the precipitations at four surrounding stations, P_t^j, $j = 1, \ldots, 4$, are available at a hourly time step for over thirty years but we use approximately ten years of data, from 1986 to 1996 (see

http://www.antony.cemagref.fr/qhan/Site%20orgeval/Page%20accueil%20francais.htm for more details on the data and the basin.). We also have daily average temperatures at this site for the same time period. Date variables serve to capture the cycles and trends in the data. Precisely, there are 16 covariates to predict the river runoff distribution: rainfall from the four precipitation stations at the previous time step, the runoff at the two previous time steps, moving averages and standard deviations with daily, weekly and monthly window widths, three date variables concerning the year, the month and the week and the daily average temperature at the previous day. Three time periods where there is no missing data are split into training and test sets. The data sets are summarized in Table 1. For this experiment, we set $Y_t = Q_{t+1}$ and $X_t = [Q_t, Q_{t-1}, P_t^1, \dots,]$ which means that given information available at time t, we model the distribution of the runoff at time $t + 1$. With the hourly data, we thus model the conditional distribution of the runoff at the next hour. In order to increase the prediction horizon to 6 and 12 hours, the hourly data are aggregated to form 6h and 12h time steps. To this end, we take the average of the runoff and the sum of the rainfall over the appropriate time period. This means that the lengths of our initial data sets in Table 1 are divided by the length of the time steps. We thus have three different models, one for each time step.

We assume that given the covariate vector X_t, the Y_t are independent and identically distributed. It is thus possible to perform model selection via five-fold cross-validation (as opposed to sequential cross-validation which is more computationally intensive, see Bishop for details [Bis95]). Model selection works as follows. The training set is divided into five subsets or folds. The conditional hybrid Pareto mixture is first trained on four of those folds

for each set of hyper-parameters considered and the performance of each trained model is evaluated on the left out fold. This process is repeated five times, so that each fold in turn was left out and that the model performance was evaluated on all the data of the training set. The hyper-parameters that gave the best performance in validation are selected. The model with the selected hyper-parameters are trained again this time on the whole training set. The generalization ability, that is the performance on unseen data, is then estimated on the test set, which is distinct from the training set. Results from the experiments on the Orgeval basin data are summarized in Table 2 for each time step (1h, 6h, 12h). The selected hyper-parameters for the penalty term, $(\lambda, \tau, \eta, \sigma)$, correspond to the prior belief that the distribution is heavy-tailed. The confidence interval is computed from the conditional quantiles of level 0.05 and 0.95, therefore, the observed runoff should fall into that interval nine times out of ten. The percentage given on the row *Confidence Interval* is the actual percentage of observed runoff on the test set which fall into the confidence interval. We can see that it is pretty close to the expected one. A mesure of goodness-of-fit is the so-called R-square given as $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, where y_i is the observed runoff, $\hat{y}_i$ is the prediction and $\bar{y}$ is the sample average. The closer R^2 is to one, the better the prediction is. The R-square is computed on the test set and the conditional median of the trained model is used to predict the runoff. We can see from the last row of Table 2 that the R-square for all time steps are very good, although the accuracy of the prediction decreases with the length of the time step. Prediction at longer time steps are understanbly more difficult. A different test set is used for the 12h time step data (the data set number 2 in Table 1) in order to leave more data for the training set. The prediction is possibly more challenging on

that time period and at least, not directly comparable with the other two models, 1h and 6h, which uses a similar test set.

The river runoff for the test period is illustrated in the left column of Figure 4, each row corresponding to one time step. The model prediction, which is the conditional median of the trained model, is plotted for each test set in the right panel of Figure 4. For all time steps, we can see that the model captured very well the dynamics of the river runoff. In the left panel of Figure 5, we have plotted the confidence intervals in light grey with quantiles of level 0.05 and 0.95 for the first 100 points of the test set. The black line is the observed runoff. Sometimes, the confidence interval is very narrow while it grows larger where the model perceives more uncertainty. We can check the effect of the tail penalty by looking at the distribution of the tail indexes of the conditional hybrid Pareto mixture on the test set. This is illustrated in the right panel of Figure 5 by an histogram. Except for a few cases in which the tail index exceeds one (which is allowed by the prior), the largest tail index values vary between 0.2 and 0.6 while most tail indexes take on values near zero. The distribution of the tail indexes is thus consistent with our prior belief.

3.2 Synthetic Data

We generate synthetic data which resemble the runoff data in the sense that there are cycles and that the tail indexes are in the same range. Let Y be a random variable distributed according to a Fréchet distribution whose parameters are functions of an input variable X.

Then the distribution function of $Y|X = x$ is given by:

$$P(Y \leq y|X = x) = \begin{cases} 0 & \text{si } y \leq \mu(x), \\ \exp\left\{-\left(\frac{y-\mu(x)}{\sigma(x)}\right)^{-1/\xi(x)}\right\} & \text{si } y > \mu(x). \end{cases}$$

The Fréchet distribution is a canonical heavy-tail distribution: the tail of most heavy-tailed distribution eventually behaves like the Fréchet tail. The input variable X is distributed according to a standard Normal distribution. We chose the following sine-shaped functional form for the dependence function $\xi(\cdot)$:

$$\xi(x) = \beta_1 + \beta_2 \sin(\gamma_1 + \gamma_2 x).$$

Since $X \sim \mathcal{N}(0, 1)$, we select the parameters of $\xi(\cdot)$ so that $\xi(X) \in [0.25, 0.5]$ with probability 0.99. The dependence function $\mu(\cdot)$ and $\sigma(\cdot)$ have a similar sine-shaped form but their parameters are chosen so that $\mu(X) \in [2, 6]$ and $\sigma(X) \in [0.5, 1]$ with probability 0.99. We generated pairs of observations (X_i, Y_i) according to this generative model. The left panel of Figure 6 illustrates the training set which is made of 2 000 such pairs of observations. The right panel shows the corresponding tail indexes. Model selection (the choice of the proper set of hyper-parameters) is performed via five-fold cross-validation on the training set. Results are presented on a test set, distinct from the training set, which consists of 10 000 pairs of observations generated according to the conditional Fréchet distribution described above.

The model selected via five-fold cross-validation for the training set of Figure 6 has eight hidden units and two mixture components. The hyper-parameters for the tail penalty are the following: $\lambda = 0.1$, $\tau = 0.45$, $\eta = 50$ and $\sigma = 0.05$. This corresponds to the shape of a prior density for heavy tails in Figure 3. The effect of the tail penalty can be seen in

the left panel of Figure 7: the histogram of the conditional tail indexes of the conditional hybrid Pareto mixture on the test set reflects the shape of the prior density. Note that less than 1% of the tail indexes are larger than 1 and are thus not shown in the Figure, this is due to the upper tail of the prior which still has some significative density in that area. For the generative model, the conditional tail indexes $\xi(X)$ vary between 0.25 and 0.5 (see the right panel of Figure 6). According to our prior belief, there should be a small subset of tail indexes from the conditional hybrid Pareto mixture which take care of modelling the upper tail and thus should take values in the same interval $[0.25, 0.5]$. The histogram of Figure 7 is consistent with this prior belief. In the right panel of Figure 7 we have plotted the test set together with the quantiles of level 0.05% and 0.95% which form a 90% confidence interval as predicted from the trained conditional hybrid Pareto mixture. Among the test set, 89% of the data points fall into the confidence interval.

In order to check how well the conditional density is learned in the upper tail, we compare three conditional quantiles of levels 0.9, 0.95 and 0.99 as computed from the generative model and the learned model. These are plotted in Figure 8: the black line is the quantile as computed from the trained conditional hybrid Pareto mixture and the light grey line is the quantile from the generative model. For the levels 0.9 and 0.95 (the top row), the two lines are almost indistinguishable from one another except for the lower and upper ends. The data density is much lower in these areas (see Figure 6) because the X variable follows a standard Normal distribution and this makes learning more difficult. The conditional quantile of level 0.99 is less well approximated. This is also due to data scarcity and shows that the model is less reliable in that case. Table 3 compares the percentage of the data in the test set which

fall below the conditional quantiles of the generative model and the trained model for the three quantile levels. The picture is pretty similar for both models. Overall, the performance of the conditional hybrid Pareto mixture with the new tail penalty proves to be satisfying.

4 Conclusion

We have propose a new stochastic model based on the conditional hybrid Pareto mixture [CB08a], in order to model the distribution of the river runoff at a future time step given rainfall observations in the hydrographic basin. This model relies on non-parametric algorithms, namely a feed-forward neural network and a mixture of distributions, from which it gains flexibility. Moreover, the component of the mixture, the hybrid Pareto, inherits the tail approximation properties of the generalized Pareto distribution which are thus transmitted to the conditional hybrid Pareto mixture. Therefore, the conditional hybrid Pareto mixture has good approximation properties, as much in the central part of the distribution as in the upper tail area.

We have introduce a penalty term in the maximum likelihood estimator in order to yield more realistic conditional tail index estimation. The penalty is based on a bimodal density which captures our prior knowledge of the distribution of the tail index. A hybrid Pareto mixture has as many tail indexes as there are components in the mixture. In the conditional case, the number of tail indexes is further multiplied by the number of data points. Our intuition is that the distribution of the tail indexes should have two modes, one around zero and one around the value of the tail index of the underlying distribution, if the latter is

heavy-tailed. Most components would be light-tailed and take care of modelling the central part of the distribution whereas few components would have a heavier tail, near the value of the tail index of the generative model, and would thus approximate the upper tail of the underlying distribution.

The conditional hybrid Pareto mixture has been trained on data from the Orgeval basin in France. Rainfall at four surrounding stations and the river runoff are available at hourly time step. These data were aggregated to obtain 6 hour and 12 hour time steps. The stochastic model was trained on three data sets, the hourly, six and 12 hour time steps. Each model can then be used to forecast the river runoff at the next hour, six or 12 hours later. Our experiments have shown that the conditional hybrid Pareto mixture is able to capture the dynamics of the basin for the three predictive time horizos. In addition, the model provides reliable confidence intervals. The tail index penalty introduces the expected distribution of the conditional tail indexes, with one mode at zero and the second mode around 0.5, more or less sharp depending on the data set.

Finally, the conditional hybrid Pareto mixture was trained on synthetic conditional data based on the Fréchet distribution. The distribution of the tail indexes is consistent with the values of the conditional tail indexes of the generative model. On the test set, 89% of the data points falls into the 90% confidence interval predicted by the model. Moreover, the trained model compares favorably with the generative model in terms of extreme quantiles.

The conditional hybrid Pareto mixture with the new penalty term has proven to be effective at modelling the rainfall-runoff process for various time steps on the Orgeval basin and more insight into the model was gain by looking at an experiment on synthetic data.

This model is very flexible and could be useful to model the rainfall-runoff process in other hydrographic basins, by using appropriate covariates.

Acknowledgments

The authors thank the following funding organizations: FQRNT, CNRS, CEA and the AssimileX and ACQWA projects.

References

C. Bishop, *Neural networks for pattern recognition*, Oxford, 1995.

P. Bernadara, D. Schertzer, E. Sauquet, I. Tchiguirinskaia, and M. Lang, *The flood probability distribution tail: how heavy is it?*, Stochastic Environmental Research and Risk Assessment **22** (2008), 107–122.

J. Carreau and Y. Bengio, *A hybrid pareto mixture for conditional asymmetric fat-tailed distributions*, IEEE Transactions on Neural Networks (2008).

———, *A hybrid pareto model for asymmetric fat-tailed data: the univariate case*, Extremes (2008).

S. G. Coles and M. J. Dixon, *Likelihood-based inference for extreme value models*, Extremes **2** (1999), no. 1, 5–23.

V. Chavez-Demoulin and A. C. Davison, *Generalized additive modelling of sample extremes*, Applied Statistics **54** (2004), 207–222.

R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the lambert w function*, Advances in Computational Mathematics **5** (1996), 329–359.

C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, *A universal approximator of convex functions applied to option pricing.*, Advances in Neural Information Processing Systems, vol. 13, 2001.

P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling extremal events*, Applications of Mathematics, Stochastic Modelling and Applied Probability, Springer, 1997.

A. Frigessi, O. Haug, and H. Rue, *A dynamic mixture model for unsupervised tail estimation without threshold selection*, Extremes **5** (2002), 219–235.

Z.-Q. Lu and L. M. Berliner, *Markov switching time series models with application to a daily runoff series*, Water Resources Research **35** (1999), no. 2, 523–534.

H. R. Maier and G. C. Dandy, *Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications*, Environmental Modelling and Software **15** (2000), 101–124.

W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in fortran: the art of scientific computing*, 2nd ed., Cambridge University Press, 1992.

M. Vrac and P. Naveau, *Stochastic downscaling of precipitation: From dry events to heavy rainfalls*, Water resources research **43** (2007).

List of Tables

List of Figures

	Time period	Hourly observations
1:	03/26/86 18:00:00 to 05/22/94 08:00:00	71 487
2:	07/22/96 15:00:00 to 08/24/01 16:00:00	44 618
3:	05/30/94 18:00:00 to 06/18/96 03:00:00	17 987

Table 1: *Three periods with no missing value in the Orgeval basin data in order of decreasing lengths.*

	Hourly	6 hours	12 hours
Training data	52 846 (1)	9 913 (1)	7 455 (1,3)
Test data	10 000 (1)	2 000 (1)	3 717 (2)
(h, m)	(4,4)	(4,8)	(4,12)
$(\lambda, \tau, \eta, \sigma)$	(0.01,0.5,50,0.1)	(0.1,0.1,50,0.2)	(1,0.1,50,0.1)
Confidence Interval	91.94%	92.1%	87.6%
R^2	0.99	0.92	0.73

Table 2: *Experiments for the Orgeval basin data, for each time step (1h, 6h, 12h) we have: the sizes of the training and test sets (data set number from Table 1), the selected number of hidden units and components* (h, m) *followed by the selected penalty hyper-parameters* $(\lambda, \tau, \eta, \sigma)$*, the percentage of the runoff in the test set which falls in the predicted 90% confidence interval and the* R^2 *of the predicted median on the test set.*

	0.9	0.95	0.99
Generative model	89.64%	94.54%	98.97%
Trained model	89.16%	94.1%	98.39%

Table 3: *Experiments with the conditional Fréchet data: percentage of the data in the test set which fall below the conditional quantiles of levels 0.9, 0.95 and 0.99 for the generative and the trained models.*

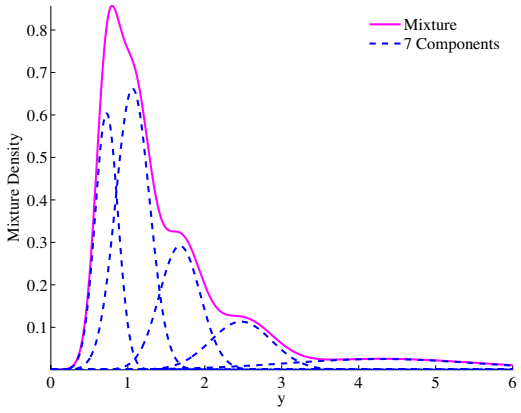

Figure 1: *Gaussian mixture density (full line) with seven components trained on heavy-tailed data. The dashed lines represent the contribution of each component to the density. Five components model the central part and the other two components contribute to the density in the upper tail.*

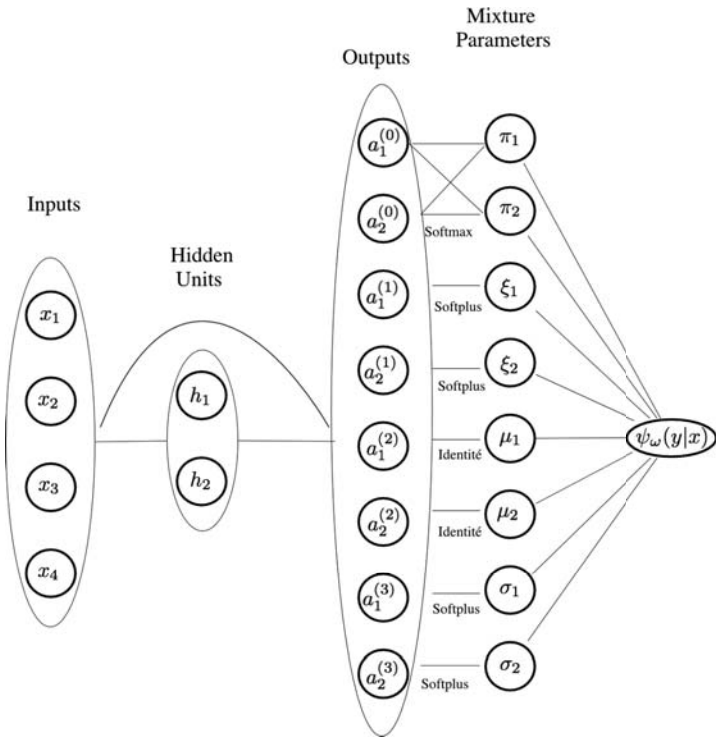

Figure 2: *Representation of a conditional mixture model with hybrid Pareto components $\psi_\omega(y|x)$. Inputs are fed to a one-layer feedforward neural network with an extra linear connection directly to the outputs. The outputs are then tranformed into the mixture parameters so as to fullfil range constraints.*

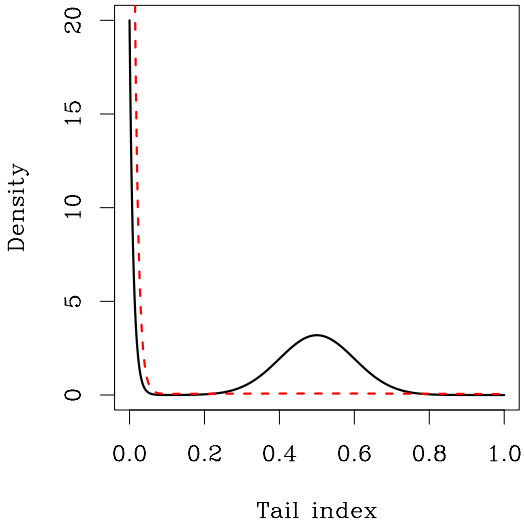

Figure 3: *The distribution in full line has one mode at zero and one mode at 0.5 while the distribution in dashed line has only significant density around zero. The former distribution reflects our prior information about how the tail indexes of a hybrid Pareto mixture should be distributed when the data is heavy-tailed and the latter distribution when the data is light-tailed.*

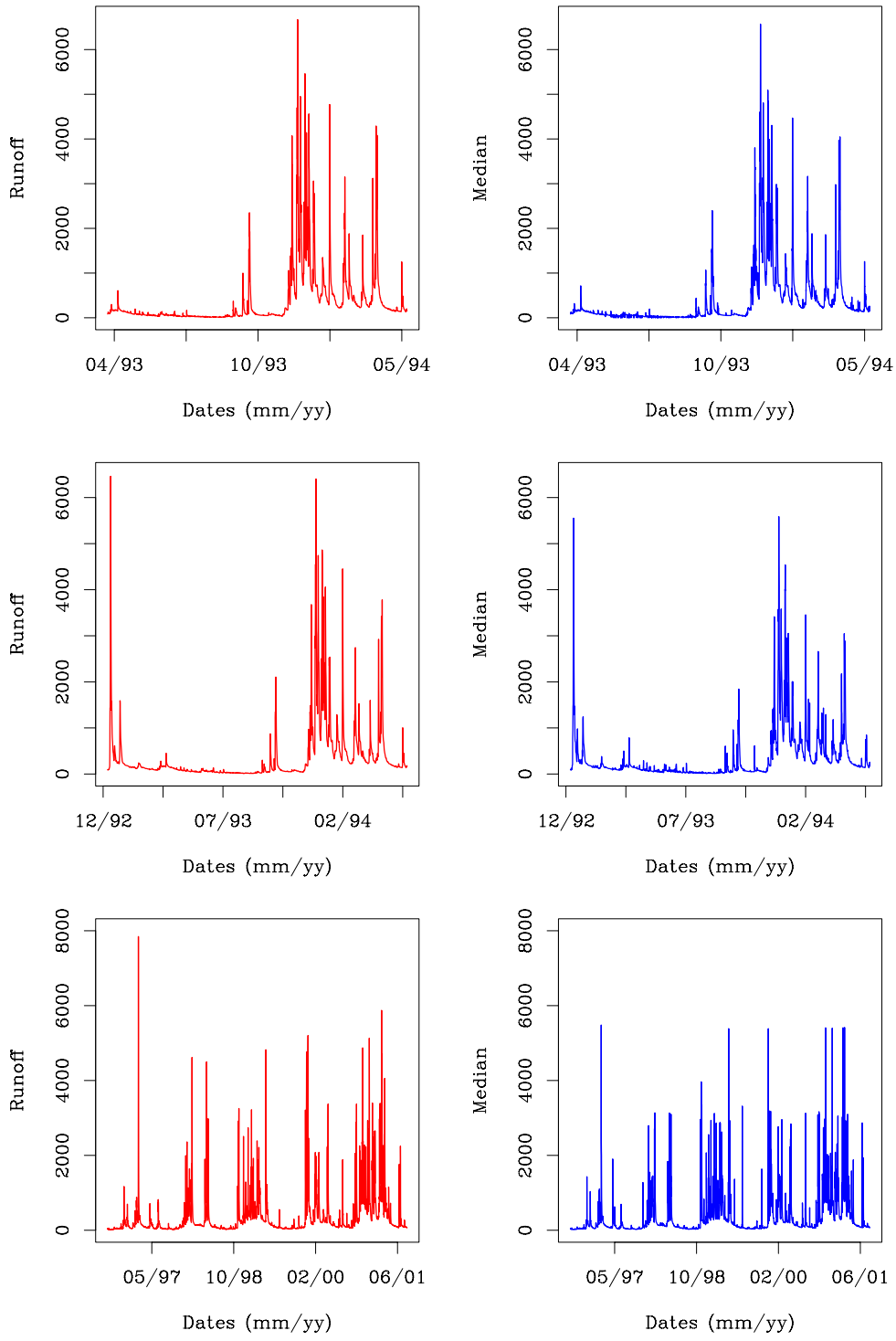

Figure 4: *Left column: observed runoff of the Avenelles sub-basin for the test period, each row corresponding to a given time step (1h, 6h and 12h). Right column: predicted median on the test set from the learned hybrid Pareto conditional mixture for the three time steps.*

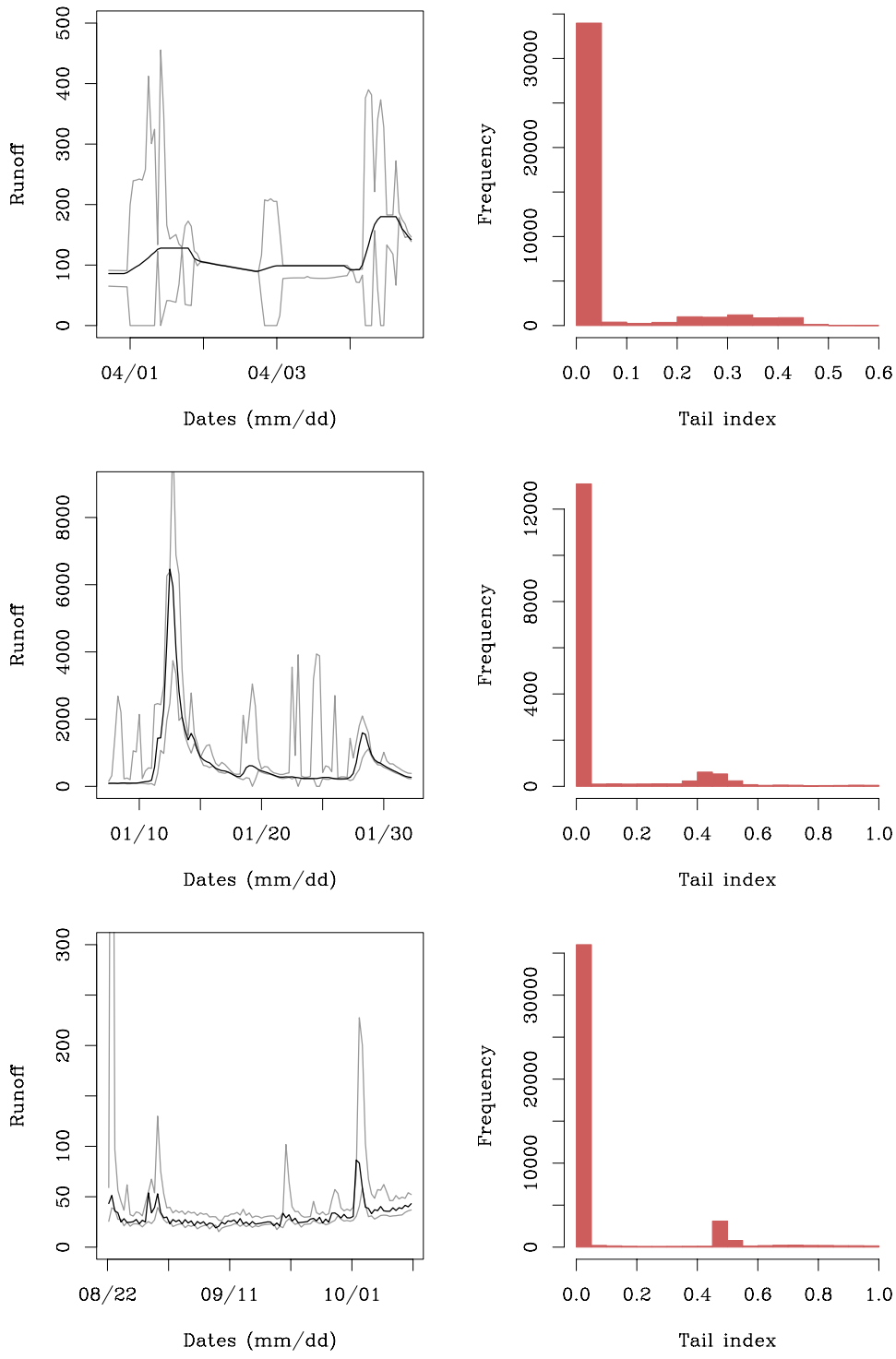

Figure 5: *Left panel: in black, the observed runoff for the first 100 points of the test set illustrated in Figure 4 together with a 90% confidence interval in light grey predicted from the conditional mixture. Right panel: histogram of the tail indexes of the conditional hybrid Pareto mixture on the test set.*

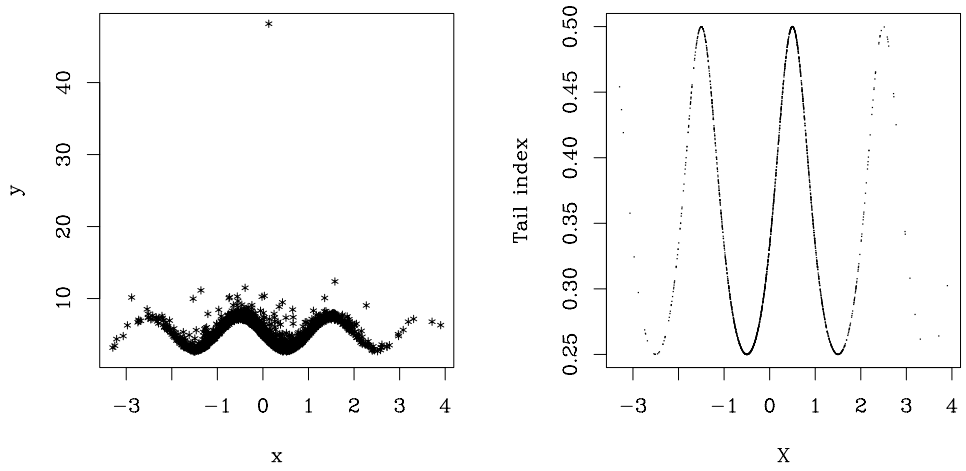

Figure 6: *Left panel: training set of 2 000 data points distributed according to the conditional Fréchet distribution with a sine-shaped functional for the dependent parameters. Right panel: the corresponding conditional tail indexes of the generative conditional Fréchet model.*

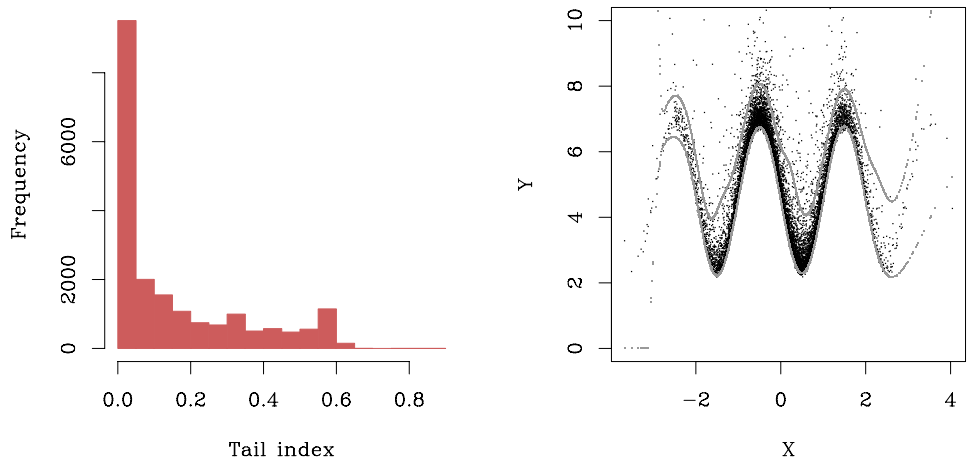

Figure 7: *Left panel: histogram of the conditional tail indexes of the trained conditional hybrid Pareto mixture on the test set. Right panel: 90% confidence interval from the trained model on the test set together with the data points (89% of the data fall into the confidence interval).*

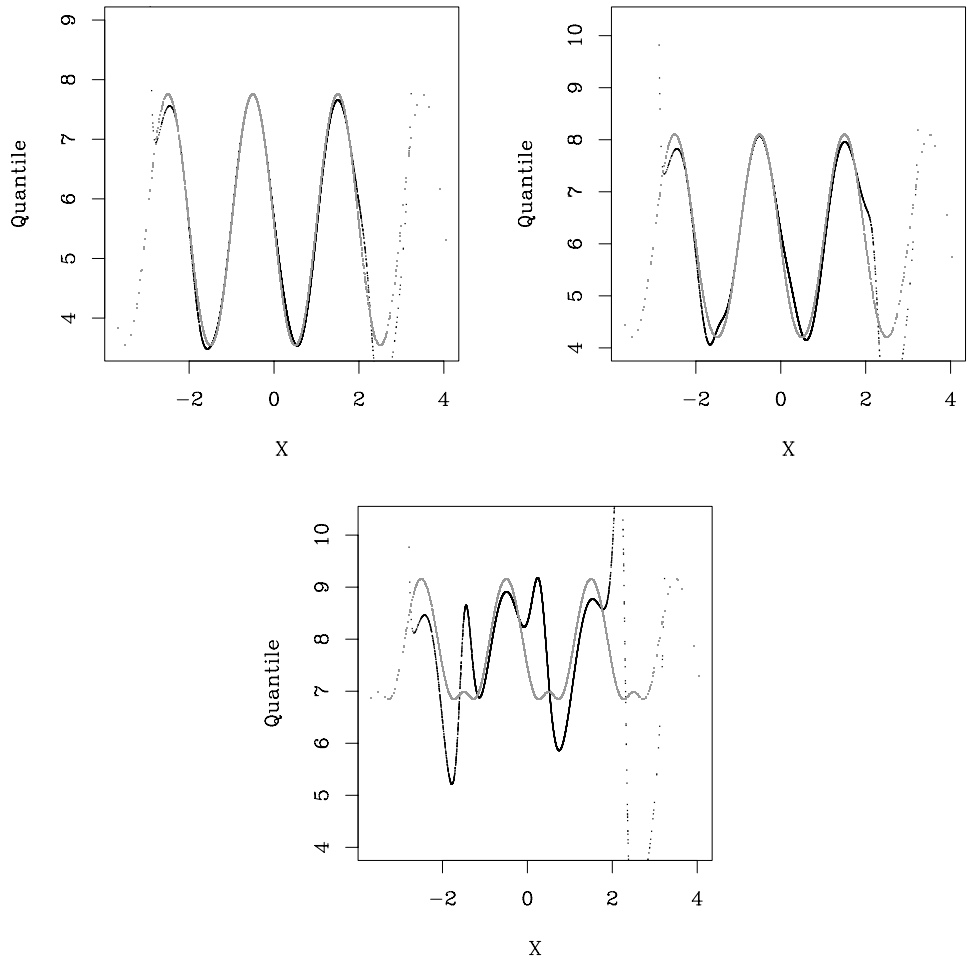

Figure 8: *Conditional quantiles of level 90%, 95% and 99% clockwise, in black, as computed from the mixture model and in light grey, from the generative conditional Fréchet model.*